BOREAL ENVIRONMENT RESEARCH 13: 195–207
ISSN 1239-6095 (print) ISSN 1797-2469 (online)

Helsinki 25 June 2008

Examining a model and assessing its performance in describing nutrient and sediment transport dynamics in a catchment in southwestern Finland

Ilona Bärlund[1)] and Teija Kirkkala[2)]

[1)] *Finnish Environment Institute, P.O. Box 140, FI-00251 Helsinki; present address: Center for Environmental Systems Research (CESR), University of Kassel, Kurt-Wolters-Strasse 3, D-34125 Kassel, Germany (e-mail: baerlund@usf.uni-kassel.de)*
[2)] *Pyhäjärvi-instituutti, Sepäntie 127, FI-27500 Kauttua, Finland (e-mail: teija.kirkkala@pyhajarvi-instituutti.fi)*

Received 2 Oct. 2006, accepted 12 Mar. 2007 (Editor in charge of this article: Raija Laiho; guest editors: Ülo Mander and Adel Shirmohammadi)

Bärlund, I. & Kirkkala, T. 2008: Examining a model and assessing its performance in describing nutrient and sediment transport dynamics in a catchment in southwestern Finland. *Boreal Env. Res.* 13: 195–207.

The Eurajoki basin, including Pyhäjärvi, was chosen as the Finnish test catchment in an EU project on benchmarking models for the Water Framework Directive due to the elevated nutrient concentrations and algal biomass production of the lake. One aim of the project was to test the suitability of models for the assessment of management options proposed to meet the surface water quality targets. Additionally, the benchmarking protocol developed to facilitate the dialogue between a modeller and a water manager in a model selection situation was tested. The catchment scale model SWAT was assessed for its applicability to analyse water and nutrient transport in Finnish environmental conditions. The results indicated that SWAT can be calibrated against measured data, especially for discharge, using a "short list" of key parameters, but further calibration is needed, especially for water quality variables. This result was supported by the attempt to validate using other monitoring points within the catchment since it revealed that the model, in the present setup, cannot reproduce observed catchment dynamics correctly. The model benchmarking guideline proved to give the process of model selection a clear structure and aided communication in a situation where the vocabulary and needs of the different parties were not established.

Introduction

Authorities responsible for water management require tools to assess the effectiveness of alternative management options since the EU Water Framework Directive (WFD) mandates Member States to develop river basin management plans for each river basin district. Effects of agricultural practices on nutrient leaching were studied in several field trials and modelling exercises in Finland (*see* e.g. Bärlund *et al*. 2007 for a summary). For catchment-scale nutrient transport two models were in focus of testing so far: INCA-N (Rankinen *et al*. 2004) and SWAT (Grizzetti *et al*. 2003, Bärlund *et al*. 2007). The first SWAT applications are promising with regard to model suitability for Finnish conditions but a detailed calibration and validation, giving an overview over a range of simulated substances, have so far not been performed. A "model" is understood here as

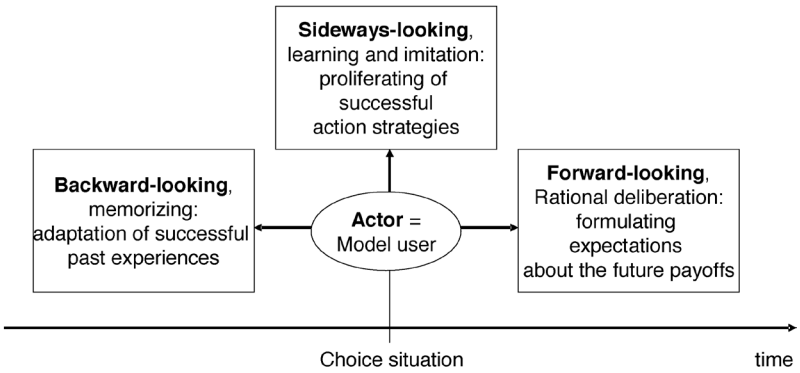

Fig. 1. The choice situation when selecting a model (modified from Heckathorn 1996).

a site-specific model established for a particular study area, including input data and parameter values (Refsgaard and Henriksen 2004).

A key question in modelling work is the selection of the "best model", if several seem to be applicable for a selected area. The practices so far applied the framework developed by Heckathorn (1996). He suggested a rational-choice based model using three ways of decision-making: backward-looking, sideways-looking and forward-looking (Fig. 1). In backward-looking decision-making the actor's behaviour is purposive and adaptive in that actions that have led to rewards in the past are more likely to be repeated, while actions that have led to losses are progressively abandoned. In modelling usually the tools that you are familiar with are the ones that perform best, i.e. the backward-looking selection could mean choosing an "institutional model" even though theoretically it would not be the most appropriate for the case-study at hand. In sideways-looking decision-making the actors compare their outcomes with those of their peers, imitating peers who do best. This applies directly to model selection since many of the models that are introduced are developed somewhere else, presented in conferences or joint projects and, if looking promising, applied to the own case-study areas. In forward-looking decision-making expectations about future events govern choices in the present. In terms of model selection this choice is made every time a model is chosen based on a manual or an external experts' recommendation promising good modelling results with a particular model code. Here "model code" is a mathematical formulation in the form of a computer program that is so generic that, without program changes, it can be used to establish a model with the same basic type of equations for different study areas (Refsgaard and Henriksen 2004).

As Axelrod (1986) notes "*… there is no need to assume that the individual is rational and understands the full strategic implications of the situation*". This also applies to model selection: thus, one aim of the EU-funded project "Benchmark models for the Water Framework Directive" (BMW) was to establish a set of criteria to assess the appropriateness of models for the use in the implementation of WFD and thus make the model selection situation more formal and transparent. These criteria developed from a set of generic questions (Saloranta *et al*. 2003) to a document that can be used as a basis for the dialogue between a modeller and a water manager (Hutchins *et al*. 2006, Kämäri *et al*. 2006). The dialogue process was supported by modelling case-studies in selected catchments. The Finnish test case in the BMW project was the catchment of Pyhäjärvi and it was based on linking models (Bärlund *et al*. 2007). First, the lake model LakeState was used for setting the targets for the loading reduction for Pyhäjärvi. Based on these results, the catchment model SWAT (ver. SWAT2000) was set up to assess the effectiveness of proposed measures to reduce agricultural and sparse settlement nutrient loading. In order to test the applicability of SWAT for this purpose, the model was applied to the Yläneenjoki catchment draining directly to Pyhäjärvi and contributing over 50% of the total P load reaching the lake (Ekholm *et al*. 1997).

The objective of this study is to describe the modelling approach using SWAT2000 for the

Finnish Yläneenjoki catchment in more detail. The work comprises several calibration steps and an attempt for validation — including the lessons learnt from them — as well as the dialogue between a modeller (I. Bärlund) and a regional water manager (T. Kirkkala) based on the benchmarking protocol developed within the BMW project.

Material and methods

The case-study area and the environmental question at stake

Pyhäjärvi, a lake situated in the municipalities of Säkylä, Eura and Yläne in southwestern Finland, is one of the most widely studied lakes in Finland. In the 1970s, the water quality of Pyhäjärvi was classified as excellent, but in the classification carried out in 2000–2003 (Vuoristo 1998 and http://www.ymparisto.fi/default.asp?contentid=130173&lan=en), the water quality was estimated as only good. The eutrophication of the lake has progressed at a rapid pace over the last few years. Pyhäjärvi is currently mesotrophic. According to studies and mathematical models, the P load to Pyhäjärvi should be reduced to almost half of the present amount in order to stop the eutrophication process and to gradually improve water quality. The major inflows to Pyhäjärvi are the rivers Yläneenjoki and Pyhäjoki, which cover 68% of the drainage basin. Of the total area 22% is cultivated; the remainder comprises forest, peatland and housing areas. Field cultivation and animal husbandry comprise 55% and 39% of the external P and N load to Pyhäjärvi, respectively. Since the drainage basin area of the lake (615 km^2) is relatively small as compared with the area of the lake itself (154 km^2), atmospheric deposition to the lake is also an important component of the external load: it makes up ca. 20% of the total P load and ca. 30% of the total N load when estimated from the bulk deposition measurements made at three stations adjacent to the lake (Ekholm *et al*. 1997).

As the Yläneenjoki catchment, 234 km^2 in area, is located on the coastal plains of southwestern Finland, the landscape ranges in altitude from 50 to 100 m a.s.l. The soils in the river valley are mainly clay and silt, whereas tills and organic soils dominate elsewhere in the catchment. Long-term (1961–1990) average annual precipitation is 630 mm of which approximately 11% falls as snow (given as the maximum water equivalent of the snow cover) (Hyvärinen *et al*. 1995). Average monthly temperature for the period of November to March, ranges anywhere between –0.5 and –6.5 °C. The warmest month is generally July when average temperature is 16.2 °C (1980–2000). Average discharge in the Yläneenjoki main channel is 2.1 $m^3\ s^{-1}$ (Mattila *et al*. 2001), which equates to an annual water yield of 242 mm (1980–1990). The highest discharges occur in the spring and late autumn: peak values vary both in spring and autumn between 5 and 38 $m^3\ s^{-1}$ (1980–2000). Groundwater contributions to stream flow are small which is reflected by a relatively low base flow index (0.40 for the period 1980–2000) and very low discharge values during low flow periods (less than 0.05 $m^3\ s^{-1}$ during 16 summers in the period 1980–2000). The main land use components in the Yläneenjoki catchment are agriculture (27% of the area), forest (48%) and peatland (21%) (Koivunen 2004). Agriculture consists of mainly cereal production and poultry husbandry. According to surveys performed in 2000–2002, 75% of the agricultural area is planted for spring cereals and 5%–10% for winter cereals (Pyykkönen *et al*. 2004). Agriculture in the Yläneenjoki catchment, as share of agricultural land of total catchment area, is intensive for Finland.

Data for only one precipitation and temperature gauge were available for the Yläneenjoki catchment (MSt, Fig. 2). The station for global radiation is located approximately 60 km outside the catchment. Regular monitoring of water quality started as early as in the 1970s in the Yläneenjoki. Monitoring of ditches and brooks entering the river or lake began in the early 1990s. The nutrient concentrations have been monitored in the Yläneenjoki by taking and analysing, in general bi-weekly, water samples and measuring the daily water flow at one point (Vanhakartano: P2 in Fig. 2). Furthermore, water quality has been monitored on a monthly basis in three additional points in the main channel and in 13 open ditches running into the Yläneenjoki since the 1990s.

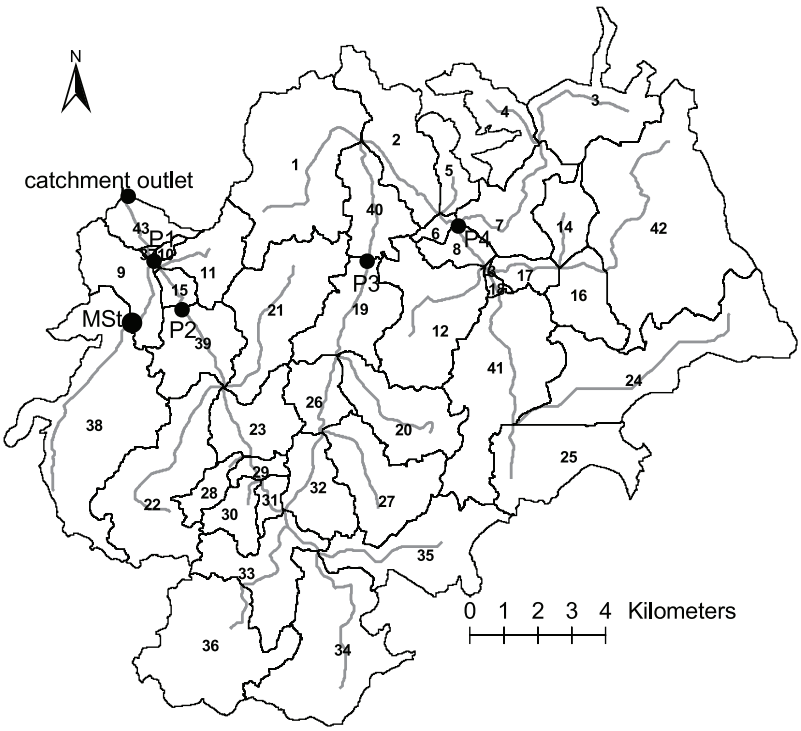

Fig. 2. SWAT setup and the location of the meteorological station (MSt) and monitoring points in the mainstream (P1–P4) in the Yläneenjoki catchment; the monitoring points in open ditches were situated from river mouth to up-streams in sub-basins 11, 38, 22, 21, 30, 33, 27, 1, 7, 14, 42 and 41.

The model and its setup

SWAT (Soil and Water Assessment Tool) is a physically based, semi-distributed river basin scale model developed to quantify the long-term impact of land management practices in large, complex watersheds (Arnold *et al*. 1998, Neitsch *et al*. 2001). It can be used to simulate water and nutrient cycles in agriculturally dominated landscapes. The catchment is generally partitioned into a number of sub-basins where the smallest unit of discretisation is a unique combination of soil and land use overlay referred to as a hydrologic response unit (HRU). SWAT is a process based model, including also empirical relationships. One objective of such a model is to assess long-term impacts of management practices. The model has been widely used but also further developed in Europe (e.g. Krysanova *et al*. 1998, Eckhardt *et al*. 2002, van Griensven *et al*. 2002). SWAT was chosen for this case study for three main reasons: its ability to simulate both P and N on catchment scale, its European wide use, and its potential to include agricultural management actions. Additionally, SWAT was evaluated by project partners against the diffuse pollution benchmark criteria developed by the BMW project and it was found to have potential with respect to the Water Framework Directive requirements (Dilks *et al*. 2003, Perrin *et al*. 2006). So in terms of Heckathorn (1996), this was a sideways-looking decision with some forward-looking elements.

For the SWAT simulations, the available data on land use and soil types had to be aggregated. The SWAT parameterisation was performed for 7 land use types: water, field, forest cuts and recently planted forest, active forest, old forest, peat bog, and sealed areas. The soil was divided into 5 general types: clay (44%), till and other coarse soils (23%), open bedrock (21%), turf (13%) and silt (0.6%), using the 25-m raster database for Finnish subsoils (depth > 25 cm) provided by the Geological Survey of Finland. Coarse soils (tills, till ridges, eskers, gravel and coarse sand), that showed a great variety but only patchwork locations within the catchment were grouped and parameterised according to the dominant type, till. The fields were parameterised to be spring barley since spring cereals are the most common crop type in the catchment. The classification of the Yläneenjoki catchment

resulted in 43 subbasins. With a threshold value of 10% for land use and for soil types the number of HRU's is 267. The parameterisation of soils and vegetation was based on measurements, expert judgement and previous field scale modelling work (i.e. using the ICECREAM model, e.g. Tattari *et al*. 2001). Clear information gaps for the Yläneenjoki data set concerned a wide range of parameters (Bärlund *et al*. 2007) where model default values are now used. Calibration took place against discharge and sediment and nutrient concentration measurements as well as calculated daily loads at Vanhakartano (P2, Fig. 2), which is situated ca. 4 km from the river mouth. This was performed for the years 1990–1994.

Results and discussion

Calibration and validation procedures

Since the hydrological calibration is described in an earlier paper (Bärlund *et al*. 2007), the emphasis here is on nutrients. The simulations were conducted without the in-stream processes option, i.e. without nutrient reactions in the river. It was discovered that the simulated annual denitrification rate in soil was unreasonably high. In the model code the denitrification limit had been set at 0.99FC (field capacity). For the clay soil that dominates the agricultural land in Yläneenjoki, this limit is clearly too low. The annual denitrification values appeared reasonable with a calibrated new limit of 1.1FC. Against a first assumption, it was later decided that crop residue removal is more close to the real situation so the parameter HVSTI (crop residues removed) was changed from 0.5 to 0.9. Except for the sum of nitrite and nitrate nitrogen (analysed and reported as NO_{23}-N), the calibration of the nutrients was mainly concentrated on point sources and initial values of the P pools in soil.

Altogether, 28 parameters were used for calibration (Table 1), which is ca. 10% of all the parameters used to run the SWAT model. The Nash-Sutcliffe Index (NSI, Nash and Sutcliffe 1970) for the different output variables varied between –263 and 0.43, the linear goodness-of-fit (R^2) values between 0.01 and 0.57 (Table 2). The best result was achieved for discharge and nutrient loads. Except for suspended sediments, the load simulation performed better than the concentration simulation. The calibration result was clearly weaker than the one previously obtained for the larger (1680 km^2) Vantaanjoki basin (Grizzetti *et al*. 2003), where NSI for the simulation of flow and total N and P loads ranged for validation from 0.43 to 0.57.

When examining time-series of simulated variables certain problems were detected:

- Discharge: no systematic errors but there still exists a discrepancy between the measured and simulated peak values during snow melt periods in winter and spring, also the low flow in summer is usually underestimated. It is possible that the use of other meteorological stations for precipitation input, even though situated outside the catchment boundaries, could improve the calibration result. The simulation of low flows is difficult due to the very low measured values (down to 0.03 m^3 s^{-1}). Now they are underestimated during certain summers by a factor of 5, especially towards the end of the summer. This has an enormous impact on the simulated concentrations of NH_4-N and PO_4-P which are strongly related to point sources.
- NH_4-N: when SWAT is run without in-stream processes, the whole NH_4-N load results from point-sources only. This means that a stable load level can be calibrated to the measurements but during low flow periods in summer the concentrations are over-estimated by a factor of 1000.
- PO_4-P: the load simulation is rather well depicted but during the low flow period the same overestimation (factor 100) of the concentration can be observed as for NH_4-N.
- NO_{23}-N, total N and P loads: the basic level is relatively well depicted but the peak values are underestimated which is mainly due to underestimation of the discharge peaks during these particular events (Fig. 3).

The overall impression is that the constant point load that is now used for scattered settlements, not connected to community waste water networks, is not working properly. It seems to be difficult to estimate the correct unit loading. The

mismatch has strong influence during low flow periods. Additionally, it is obviously not enough to base the calibration on a limited number of catchment or subbasin wide parameters (Table 1) but the singular hydrological response units (HRU's) and subbasins have to be thoroughly examined for their output and the in-stream processes have to be included in the calibration. It has to be noted, though, that not all of the ca. 300 input parameters of SWAT are relevant in this context since neither pesticides, bacteria nor metals are considered and no ponds or wetlands are included.

Even though the calibration result showed

Table 1. SWAT parameters used to calibrate discharge and nutrients at Vanhakartano.* = values are multiplicators.

No.	Parameter	Starting value	End value	Calibrated against
1	Snowfall temperature SFTMP	–3.2	–0.2	discharge
2	Snow melt base temperature SMTMP	–0.3	–0.1	discharge
3	Minimum snow water content that corresponds to 100% snow cover SNOCOVMX	1	10	discharge
4	Surface runoff lag coefficient SURLAG	4	1	discharge
5	Groundwater delay time GW_DELAY	31	25	discharge
6	Deep aquifer percolation fraction RCHRG_DP	0	0.1	discharge
7	Maximum canopy storage CANMX (forest)	5 & 10	50 & 70	discharge
8	Minimum (base) temperature for plant growth T_BASE (forest)	2, 5	0	discharge
9	Maximum potential leaf area index BLAI (forest)	5	9	discharge
10	Total number of heat units or growing degree days needed to bring plant to maturity PHU (forest)	2000, 2500	3500	discharge
11	Manning's "n" value for overland flow OV_N (agricultural land)	0.04	0.19	discharge
12	Manning's "n" value CH_N (tributaries)	0.05	0.08	discharge
13	Drain tile lag time GDRAIN	12	48	discharge
14	Peak rate adjustment factor for sediment routing in the main channel PRF	0.8	1	sediment conc.
15	Initial SCS runoff curve number fro moisture condition II CN2 (forest on clay, silt and turf)	78, 77, 70	55	discharge, NO_{23}-N
16	Initial SCS runoff curve number fro moisture condition II CN2 (forest on moraine, open rock)	25	34	discharge, NO_{23}-N
17	Initial SCS runoff curve number fro moisture condition II CN2 (agricultural land on clay)	83	70	discharge, NO_{23}-N
18	Initial SCS runoff curve number fro moisture condition II CN2 (agricultural land on clay, tillage)	83, 89	75	discharge, NO_{23}-N
19	Clay content CLAY (agric. subsoil, clay)	74	55	discharge, NO_{23}-N
20	Moist bulk density SOL_BD (agric. subsoil, clay)	1.1	1.3	discharge, NO_{23}-N
21	Available water capacity of the soil layer SOL_AWC (agric. subsoil, clay)	0.25	0.17	discharge, NO_{23}-N
22	Clay content CLAY (agric. subsoil & forest topsoil, turf)	33	3	discharge, NO_{23}-N
23	Available water capacity of the soil layer SOL_AWC (agric. subsoil & forest topsoil, turf)	0.60	0.50	discharge, NO_{23}-N
24	Nitrate percolation coefficient NPERCO	0.2	0.9	NO_{23}-N
25	Average daily mineral P loading MINPCNST (point sources)	1*	0.25*	PO_4-P
26	Initial soluble P concentration in soil layer SOL_SOLP (moraine, clay)	30, 40	20, 30	PO_4-P
27	Average daily organic P loading ORGPCNST (point sources)	1*	0.25*	P_{tot}
28	Initial organic P concentration in soil layer SOL_ORGP (all soils)	465	207 (calc. intern.)	P_{tot}

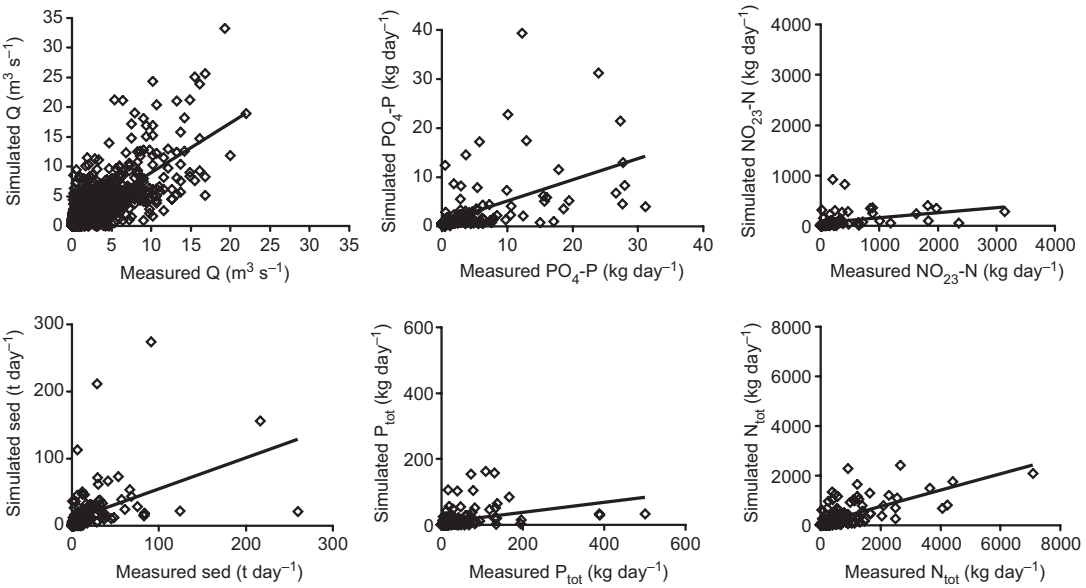

Fig. 3. SWAT calibration result as relationship of simulated daily values against measured discharge (*Q*) and calculated daily loads of phosphate phosphorus (PO_4-P), sum of nitrite and nitrate nitrogen (NO_{23}-N), suspended sediments (sed), total phosphorus (P_{tot}) and total nitrogen (N_{tot}) for the period 1990–1994 at Vanhakartano (P2 in Fig. 2).

clear weaknesses, and further calibration thus is required, the model performance analysis was continued since the overall graphical match between the measured and simulated timelines was acceptable to the water manager of this exercise. As a second step, a validation of this SWAT setup was attempted against data at the same Vanhakartano location for the period 1995–1999. The intention was to use the model run with unchanged parameter set against independent data series as a further analysis of the model performance, not as a proof of overall validity. With the exception of suspended sediment load and concentration, the validation performance was poorer than the calibration result (Table 3). This result is confirmed in many modelling papers (e.g. Refsgaard and Henriksen 2004). The comparison of NO_{23}-N simulation performance in two different years 1996 and 1999 shows, however, that the judgement of model performance is difficult for sparse measurement data (Fig. 4). In 1996 the simulated result followed well the measured concentration level. In early autumn 1999, however, the NO_{23}-N concentrations were overestimated by the model by a factor of 10. Also the measurements show a rise in NO_{23}-N concentration from October to December 1999, probably as a reaction to elevated mineralisation potential after an exceptionally dry summer, but this process started too early in the autumn and

Table 2. The evaluation of the calibration result at Vanhakartano (P2 = sub-basin 39) for the period 1990–1994. NSI = Nash-Sutcliffe index, R^2 = linear goodness-of-fit, *n* = number of measurement-simulation pairs.

Variable	NSI	R^2	*n*	Variable	NSI	R^2	*n*
Discharge	0.43	0.57	1826	N_{tot} load	0.32	0.46	180
Sediment load	–0.11	0.21	172	N_{tot} conc.	–2.2	0.01	180
Sediment conc.	0.01	0.20	172	PO_4-P load	0.15	0.29	171
NH_4-N load	0.01	0.02	124	PO_4-P conc.	–9.3	0.03	171
NH_4-N conc.	–263	0.02	124	P_{tot} load	0.01	0.13	191
NO_{23}-N load	0.16	0.14	95	P_{tot} conc.	–2.0	0.07	191
NO_{23}-N conc.	–0.11	0.22	95				

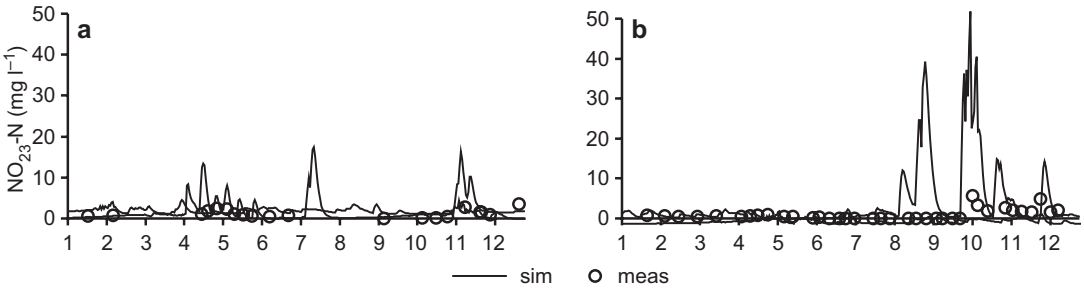

Fig. 4. Simulated and measured NO_{23}-N concentrations in (**a**) 1996 and (**b**) 1999 at Vanhakartano P2 in the validation period 1995–1999.

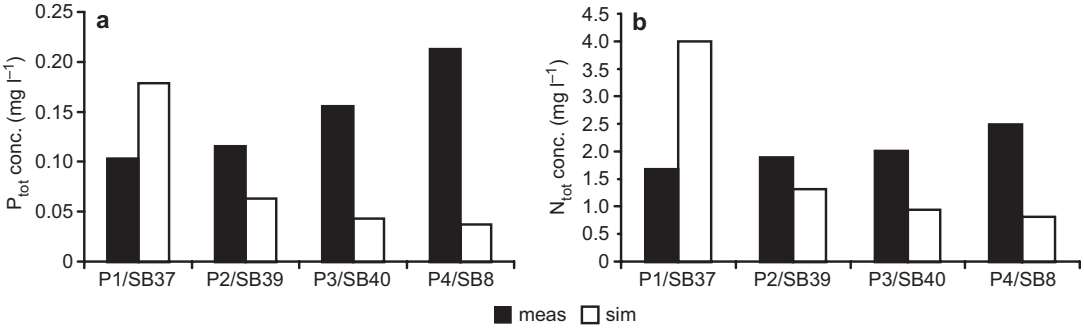

Fig. 5. Simulated and measured average annual concentrations of (**a**) P_{tot} and (**b**) N_{tot} at four measurement points P1–P4 in the main stream for the period 1991–1994.

was heavily overestimated by the model. Additional sources for the discrepancy between the measured and simulated concentrations during autumn 1999 could be the underestimation of the flow peaks and lack of in-stream denitrification in the model.

A second validation was performed concerning the average concentrations of total nutrients along the main stream (points P1–P4, Fig. 2). This analysis revealed further major problems in the present model setup to describe catchment dynamics (Fig. 5). The measured average concentrations for the years 1991–1994 indicated a rise from the point closest to the river mouth (P1) to the agriculturally intensive upper parts of the catchment (P4) — the simulation results showed just the opposite. This result indicates that, in the present model setup, the main catchment element affecting the simulation result are the processes in the stream like dilution, not the loading from land reflecting land use. This is not what is commonly understood and not

Table 3. The evaluation of the validation result at Vanhakartano (P2 = sub-basin 39) for the period 1995–1999, additionally 1985–1989 for discharge (Q) (NSI = Nash-Sutcliffe index, R^2 = linear goodness-of fit, n = number of measurement-simulation pairs).

Variable	NSI	R^2	n	Variable	NSI	R^2	n
Q (1995–1999)	0.18	0.32	1826	NO_{23}-N conc.	–53	0.03	157
Q (1985–1989)	0.24	0.42	1826	N_{tot} load	–2.7	0.33	191
Sediment load	–0.18	0.21	191	N_{tot} conc.	–34	0.07	191
Sediment conc.	0.10	0.11	191	PO_4-P load	–0.39	0.10	181
NH_4-N load	0.0	0.00	155	PO_4-P conc.	–17	0.01	181
NH_4-N conc.	–401	0.01	155	P_{tot} load	–3.5	0.05	189
NO_{23}-N load	–8.3	0.37	157	P_{tot} conc.	–21	0.00	189

what the measurements show. The agricultural land is now discretised as being spring barley with a moderate inorganic fertilisation practise so one possibility to improve the model setup is to include eventual hot spots in form of HRU's receiving organic fertilisation. Another possibility would be to examine what the effect of the downstream forested areas is, as the result might be due to overestimated loading from forested HRU's. The effect of calibration can be seen as the best fit found at the Vanhakartano calibration point P2 (Fig. 5).

The final validation test was performed using measured and simulated concentrations from 12 monitored open ditches running into the Yläneenjoki (Fig. 2). Average concentrations from the period 1991–1994 were compared against the share of agricultural land in each of the sub-catchments contributing to the monitoring point. The discretisation of the sub-catchments in the SWAT model and in the monitoring programme was not identical but the linear goodness-of-fit (R^2) value of 0.89 indicated a good agreement (Fig. 6a). Both the measured average NO_{23}-N concentration and measured average suspended sediment concentration showed a strong linear correlation (R^2 = 0.92 and 0.73, respectively) with the share of agricultural land in the sub-catchment. This indicates a strong relationship of the NO_{23}-N concentration to N fertilisation and of the suspended sediment concentration to the open agricultural field land susceptible to erosion. The modelled performance for NO_{23}-N (Fig. 6b) strengthened the impression from Fig. 4 that this variable, in principle, can be well depicted by SWAT, if the calibration is improved. This was shown by the similarly strong linear correlation but with so far too low values for the sub-catchments with high share of agricultural land. The opposite was shown for suspended sediments (Fig. 6c). Here the good linear correlation from the measurements could not be reproduced. It seemed that irrespective of the share of agricultural land the concentration level of ca. 20 mg l^{-1} is simulated on average. Since all the 12 open ditches are of similar magnitude, the suspicion from the previous validation experiment in the main channel that the concentration of certain variables in the surface waters is not governed by the transport from land but the channel processes

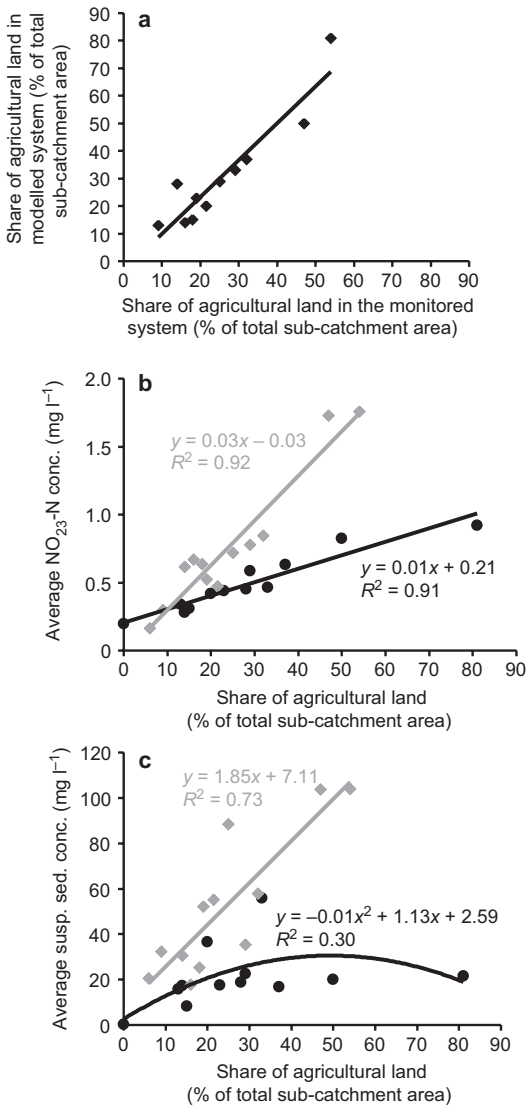

Fig. 6. The relationship between (**a**) the share of agricultural land in the monitored and modelled sub-catchments, (**b**) the monitored/modelled share of agricultural land and the monitored/modelled average NO_{23}-N concentrations, and (**c**) the monitored/modelled share of agricultural land and the monitored/modelled average suspended sediment concentrations in 12 sub-catchments for the period 1991–1994. Measured in gray, simulated in black.

themselves is strengthened. This is probably also a calibration problem but additionally it should be investigated if the average slope derived from the digital elevation model for the whole sub-catchment is sufficient to describe the overland erosion process correctly — especially in an area like the

Yläneenjoki catchment, which is characterised by small scale variability of the landscape.

The dialogue between the modeller and the water manager

The model benchmarking protocol that was created within the BMW project consists of a set of 23 questions (Hutchins *et al*. 2006 and Appendix) for the water manager and modeller to consider in a joint model selection session. The issues are divided into four sections after each of which a GO or NO GO decision has to be made:

1. Definition of the management and modelling tasks.
   - GO/NO GO: Is modelling needed?
2. Model functionality and data.
   - GO/NO GO: Is the model code suitable for this task?
   - GO/NO GO: Can the model be used for this application?
3. Model performance assessment
   - GO/NO GO: Does the model perform in an acceptable way in this application?
4. A posteriori review.
   - GO/NO GO: Can the model be used for the management tasks at hand?

Benchmarking based on the Yläneenjoki case showed a clear GO for the two first steps even though certain reservations were noted:

- Model structure: HRU's are defined as a share of a sub-basin but they do not have coordinates (e.g. no distance to the water course), which is vital in the planning of management actions like buffer strips; also some processes which might be of importance are missing like denitrification in streams and the NH_4-N pool in soil.
- Empirical equations: SWAT does not route water using mass conservation-based continuity equations but use the USDA SCS runoff curve number method to compute runoff volumes. In addition, SWAT uses an empirical procedure to route water through channels (e.g. Borah and Bera 2003). These empirical approaches have been criticised for their lack of transferability from one part of the world to another without additional calibration.
- Data availability: there is a discrepancy between the GIS land use information available (one class for agricultural land) and the agricultural management practices (cereals, grass, root crops, etc.); there was only one meteorological station within the catchment even though several might be needed to depict e.g. snow melt events better; SWAT contains a large number of input parameters of which many are empirical and as such have no true chemical or physical meaning.
- Examples of regional model use: sufficient experience of model performance has not yet been gathered in northern environmental conditions. Many water managers are not convinced by scientific arguments and cannot evaluate the model code, so the acceptance relates to successful applications of the model to national conditions.

The model performance assessment (step 3) was not completed during the project due to the time consuming calibration and validation effort of a complex model like SWAT when looking simultaneously at hydrology, erosion, and suspended sediments as well as leaching and reactions of inorganic and organic fractions of both N and P. A simple filter strip exercise performed with this preliminary calibration (Bärlund *et al*. 2007) convinced the water manager, however, of continuing the time-consuming calibration and validation exercise, since it appears that once the SWAT application has gained sufficient confidence it actually can be utilised to demonstrate effects of alternative management actions and thus support decision making.

One of the most important remaining questions from this benchmarking exercise is whether a sufficient confidence in model performance exists. This could be, according to the water manager in this exercise, several credible model applications or a comparison with the level of uncertainty in the available field observations (Refsgaard and Henriksen 2004), as no universal accuracy criteria can be established. Further, the role of the water manager is not only as a discussion partner but also as a source of "soft data", qualitative knowledge which cannot be used

directly as exact numbers but can be applied in evaluating model performance and parameter value acceptability (Seibert and McDonnell 2002). One may also consider the stochastic nature of the input parameter values and determine the impact of such variability on model results (Shirmohammadi *et al*. 2006).

Conclusions and outlook

The approach to assess SWAT model performance in the Finnish Yläneenjoki catchment at the calibration point Vanhakartano revealed that the model can be calibrated to discharge and nutrient loads using a limited parameter set of ca. 30 input parameters, but especially for water quality variables further calibration is required. The validation attempt at the same Vanhakartano point indicated that the calibration performance directly translates into validation performance. With regard to this part of the evaluation, the present version of SWAT would be acceptable to the water manager as a tool to be used for management actions like installation of buffer strips. The further validation work within the catchment showed, however, that the calibration and validation — even using a split-sample test — to one point is not enough to provide understanding of the dynamics of such a complex model like SWAT. This confirms the point made by Refsgaard and Henriksen (2004) that establishment of validation test schemes for the situations, where the split-sample test is not sufficient, is an area, where limited work has been carried out so far and to which more attention thus should be paid.

Three options for continued work remain: (1) improve calibration using sub-basin and HRU level information more efficiently and pay attention to the in-stream processes; (2) improve the model by changing e.g. snow accumulation and melting routines and the description of forested areas on organic soils; (3) choose another model. Given the situation that the availability of models which fulfil the requirements of simulating both P and N on catchment scale and including agricultural management actions in Finland is limited and that the simple exercises performed so far using the present setup for buffer strip efficiency demonstrations is valued by the regional water manager, a further improvement of the calibration (N) and a consideration of model improvements (erosion and suspended sediment transport) is recommended. It is evident, however, that SWAT is a very data intensive model and its applicability is thus restricted — especially if transferability of parameter sets from one catchment to another cannot be proven. The appropriate use of a model like SWAT is time consuming and requires an experienced user. This is a further aspect that has to be considered when planning to use the model for practical water management issues.

This work highlights the importance of a decision aid like the BMW benchmarking protocol to be used in discussions between a modeller and a water manager. The sideways-looking model selection adopted in the beginning of this project clearly led to an underestimation of the time needed to setup, calibrate and validate the model and thus the whole modelling task could not be completed within the permitted time-frame. In reality there is seldom a selection of alternative models or modellers available to be switched to if one model shows a NO GO at a benchmarking situation. Saloranta (2006) gives an example of benchmarking four lake models for a Norwegian case study where two of the models were actually screened out before the actual evaluation using the benchmarking criteria. In any case, the benchmarking protocol gives the process of model selection a clear structure and aids communication in a situation where the vocabulary and needs of the different parties are not yet established. It may thus not just be the model choice that would gain from a benchmarking process but also the data-sets used for model evaluation. Going through a similar benchmarking scrutiny, the data-sets would have a similar backbone of criteria to be used arguing for or against a model choice.

Acknowledgements: The financial support of the "Benchmark models for the Water Framework Directive" project through the European Commission's 5th Framework Programme (contract EVK1-CT-2001-00093) is gratefully acknowledged. The authors wish also to thank colleagues in the Research Programme for Environmental Policy (PTO) of the Finnish Environment Institute for making us aware of Heckathorn's approach and for commenting a natural science modelling manuscript from a social science perspective.

References

Arnold J.G., Srinivasan R., Muttiah R.S. & Williams J.R. 1998. Large area hydrologic modelling and Assessment part I: model development. *Journal of American Water Resources Association* 34: 73–89.

Axelrod R. 1986. An evolutionary approach to norms. *The American Political Science Review* 80: 1095–1111.

Borah D.K. & Bera M. 2003. Watershed-scale hydrologic and nonpoint-source pollution models: review of mathematical bases. *Transactions of the ASAE* 46: 1553–1566.

Bärlund I., Kirkkala T., Malve O. & Kämäri J. 2007. Assessing SWAT model performance in the evaluation of management actions for the implementation of the Water Framework Directive in a Finnish Catchment. *Environmental Modelling & Software* 22: 719–724.

Dilks C.F., Dunn S.M. & Ferrier R.C. 2003. Benchmarking models for the Water Framework Directive: evaluation of SWAT for use in the Ythan catchment, UK. In: Srinivasan S., Jacobs J.H. & Jensen R. (eds.), *Condensed abstracts of the 2nd International SWAT Conference, 1–4.7.2003, Bari, Italy*, TWRI Technical Report 266, pp. 202–207.

Eckhardt K., Haverkamp S., Fohrer N. & Frede H.-G. 2002. SWAT-G, a version of SWAT99.2 modified for application to low mountain range catchments. *Physics and Chemistry of the Earth* 27: 641–644.

Ekholm P., Malve O. & Kirkkala T. 1997. Internal and external loading as regulators of nutrient concentrations in the agriculturally loaded Lake Pyhäjärvi (southwest Finland). *Hydrobiologia* 345: 3–14.

Grizzetti B., Bouraoui F., Granlund K., Rekolainen S. & Bidoglio G. 2003. Modelling diffuse emission and retention of nutrients in the Vantaanjoki watershed (Finland) using the SWAT model. *Ecological Modelling* 169: 25–38.

Heckathorn D.D. 1996. The dynamics and dilemmas of collective action. *American Sociological Review* 61: 250–277.

Hutchins M.G., Urama K., Penning E., Icke J., Dilks C., Bakken T.H., Perrin C., Saloranta T., Candela L. & Kämäri J. 2006. The model evaluation tool: guidance for applying benchmark criteria for models to be used in river basin management. *Archiv für Hydrobiologie*, *Supplement volume* 161, *Large Rivers* 17: 23–48.

Hyvärinen V., Solantie R., Aitamurto S. & Drebs A. 1995. *Water balance in Finnish drainage basins during 1961–1990*. Publications of Water and Environment Administration A220. [In Finnish with English abstract]

Koivunen S. (ed.) 2004. *Yläneenjoki — vesiensuojelu ja virkistyskäyttö*. Pyhäjärvi-instituutin julkaisuja, sarja B, no. 11.

Krysanova V., Müller-Wohlfeil D.-I. & Becker A. 1998. Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds. *Ecological Modelling* 106: 261–289.

Kämäri J., Boorman D., Icke J., Perrin C., Candela L., Elorza F., Ferrier R., Bakken T.H. & Hutchins M. 2006. Process for benchmarking models: dialogue between water managers and modellers. *Archiv für Hydrobiologie*, *Supplement volume* 161, *Large Rivers* 17: 3–21.

Mattila H., Kirkkala T., Salomaa E., Sarvala J. & Haliseva-Soila M. (eds.) 2001. *Pyhäjärvi*. Pyhäjärvi-instituutin julkaisuja 26. Pyhäjärven suojelurahasto.

Nash J.E. & Sutcliffe J.V. 1970. River flow forecasting through conceptual models. Part I: A discussion of principles. *Journal of Hydrology* 27: 282–290.

Neitsch S.L., Arnold J.G., Kiniry J.R. & Williams J.R. 2001. *Soil and water assessment tool — theoretical documentation — Version 2000*. Blackland Research Center, Agricultural Research Service, Texas, USA.

Perrin C., Dilks C., Bärlund I., Payan J.L. & Andréassian V. 2006. Use of simple rainfall-runoff models as a baseline for the benchmarking of the hydrological component of complex catchment models. *Archiv für Hydrobiologie*, *Supplement volume* 161, *Large Rivers* 17: 75–96.

Pyykkönen S., Grönroos J., Rankinen K., Laitinen P., Karhu E. & Granlund K. 2004. *Cultivation measures in 2000–2003 and their effects to the nutrient runoff to the waters in the farms committed to the Agri-Environmental Programme*. The Finnish Environment 711. [In Finnish with English abstract].

Rankinen K., Granlund K. & Lepistö A. 2004. Integrated nitrogen and flow modelling (INCA) in a boreal river basin dominated by forestry: scenarios of environmental change. *Water, Air and Soil Pollution, Focus* 4: 161–174.

Refsgaard J.C. & Henriksen H.J. 2004. Modelling guidelines — terminology and guiding principles. *Advances in Water Resources* 27: 71–82.

Saloranta T.M., Kämäri J., Rekolainen S. & Malve O. 2003. Benchmark criteria: a tool for selecting appropriate models in the field of water management. *Environmental Management* 32: 322–333.

Saloranta T.M. 2006. Highlighting the model code selection and application process in policy-relevant water quality modelling. *Ecological Modelling* 194: 316–327.

Seibert J. & McDonnell J.J. 2002. On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. *Water Resources Research* 38(11) 1241, doi:10.1029/2001WR000978.

Shirmohammadi A., Chaubey I., Harmel R.D., Bosch D.D., Muñoz-Carpena R, Dharmasri C., Sexton A, Arabi M., Wolfe M.L., Frankenberger J., Graff C. & Sohrabi T.M. 2006. Uncertainty in TMDL models. *Transactions of ASAE* 494: 1033–1049.

Tattari S., Bärlund I., Rekolainen S., Posch M., Siimes K., Tuhkanen H.-R. & Yli-Halla M. 2001. Modelling sediment yield and phosphorus transport in Finnish clayey soils. *Transactions of ASAE* 44: 297–307.

Van Griensven A., Francos A. & Bauwens W. 2002. Sensitivity analysis and autocalibration of an integral dynamic model for river water quality. *Water Science and Technology* 45: 321–328.

Vuoristo, H. 1998. Water quality classification of Finnish inland waters. *European Water Management* 1: 35–41.

Appendix

BMW Benchmarking protocol (for details *see* Hutchins *et al*. 2006, Kämäri *et al*. 2006): the protocol includes in addition to the questions themselves general guidance and explanations as well as considerations to be reflected upon from the modeller's point of view and from the water manager's side, respectively.

Section 1

1. What is the problem?
2. What are the main causes of the problem?
   - 2.1 Make a conceptual model of the problem.
   - 2.2. Define the broad management objective(s).
3. What measures may be implemented to achieve the management objective(s) stated in 2.2?
4. GO/NO GO: Is any modelling approach appropriate?

Section 2

5. Does the model output meet the requirements of the management task(s)?
6. Does the model include the processes and components relevant to the management task?
7. Does the temporal and spatial span / resolution of the model code correspond to the management task?
8. Are the data required for the implementation of the model available?
9. GO/NO GO: The selected model code is potentially suitable for this management task.
10. Is there sufficient scientific and stakeholder acceptance of the model code?
11. Is there sufficient guidance to aid model application?
12. Has the model code been sufficiently tested?
13. Does the model code have version control?
14. Is the user interface appropriate for the application and user?
15. How identifiable are the model parameters?
16. Is there sufficient understanding of the model's uncertainty and sensitivity?
17. Is the model code sufficiently flexible for adaptation, improvements and linking?
18. GO/NO GO: Is the model code suitable for this application?

Section 3

19. Is the model application response sufficiently consistent with your understanding of the behaviour of the natural system?
20. Is the assessment of the model application performance satisfactory?
21. Is the uncertainty in the output of the model application satisfactory addressed?

Section 4

22. How useful was the model application for informing the management?
23. What are the recommendations that follow from the modelling study?